SUPPLEMENTARY INFORMATION

Genomes of all known members of a *Plasmodium* subgenus reveal paths to virulent human malaria

Authors: Thomas D. Otto[1,†,#,*], Aude Gilabert[2, †], Thomas Crellen[1,3], Ulrike Böhme[1], Céline Arnathau[2], Mandy Sanders[1], Samuel O. Oyola[1,4], Alain Prince Okouga[5], Larson Boundenga[5], Eric Willaume[6], Barthélémy Ngoubangoye[5], Nancy Diamella Moukodoum[5], Christophe Paupy[2], Patrick Durand[2], Virginie Rougeron[2,5], Benjamin Ollomo[5], François Renaud[2], Chris Newbold[1,7], Matthew Berriman[1,*] & Franck Prugnolle[2,5,*]

Affiliations:

[1] Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton CB10 1SA, United Kingdom

[2] Laboratoire MIVEGEC, UMR 5290-224 CNRS 5290-IRD 224-UM, Montpellier, France

[3] Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, United Kingdom

[4] International Livestock Research Institute, Box 30709, Nairobi, Kenya (current address)

[5] Centre International de Recherches Médicales de Franceville, Franceville, Gabon

[6] Sodepal, Parc of la Lékédi, Bakoumba, Gabon

[7] Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, United Kingdom

Current Address: Centre of Immunobiology, Institute of Infection, Immunity & Inflammation, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, United Kingdom

[*]Correspondence to: Thomas D. Otto (ThomasDan.Otto@glasgow.ac.uk), Matthew Berriman (mb4@sanger.ac.uk) or Franck Prugnolle (franck.prugnolle@ird.fr)

Contents

Supplementary Note 1: Dating and population size estimates 3

Estimation of single nucleotide substitution per year for different generation times 3

1. Assuming 60-day generation time 4

2. Assuming 180-day generation time 4

In-vivo data 5

Conserved regions around the PfCRT in five clinical *P. falciparum* isolates. 5

Coalescent models 6

Multiple Sequentially Markovian Coalescent 6
Estimation of population size 7
Dating of *eba-175* dimorphism 7
Supplementary Note 2: Evolution of core genes 9
Within-species polymorphism 9
Interspecific gene transfer 9
Genome-wide test of convergent evolution 10
Supplementary Note 3: Gene family analyses 11
Differences in gene families 11
Generation of similarity matrices 11
The Rifin and Stevor proteins 12
Meme-Motif analysis for Stevor proteins 12
var gene analysis 12
Supplementary Figures 14
List of Supplementary Tables 1-9 31

Supplementary Note 1: Dating and population size estimates

A major focus of this study has been to understand the population history of the *Laverania* species and in particular the timing of the divergence events as well as the variation of population size. We used two methods: (1) a Bayesian coalescence model, G-PhoCS[1] to estimate the timing of species divergence (Fig. 1a) and (2) the Multiple Sequentially Markovian Coalescent (MSMC)[2] method to provide a high resolution estimate of changes of N_e through time, specifically to look for a bottleneck that would explain the low diversity in the *P. falciparum* population (Fig. 1b). To scale the population genetic parameters inferred from these models to 'real time' and N_e, we used a per-base mutation rate of 3.78×10^{-10} (for 4 mitotic events in the red blood cycle)[3].

Estimation of single nucleotide substitution per year for different generation times

The total number of mitoses per generation was calculated based on different assumptions about total generation time (time to complete the full life cycle), time to complete different stages of the life cycle and number of mitoses per stage.

Data from previous studies:

- Development in mosquito takes 10–22 days and involves 10–12 mitoses[4]
- Development in liver takes 5–7 days and involves 15 mitoses[5]
- Gametocyte development takes 12 days and involves 3 mitoses[6].
- Intra-erythrocytic development involves 2 mitoses per day (undergoes three to four rounds of DNA synthesis, mitosis, and nuclear division to produce a syncytial schizont with 16 to 22 nuclei)[6]

We have assumed that generation times can be within the range 60 –180 days[7,8].

1. Assuming 60-day generation time

	Days	Min. mitoses	Max. mitoses	Days	Min. mitoses	Max. mitoses
Oocyst to salivary gland	10	10	12	22	10	12
Liver	5	15	15	7	15	15
Gametocytes	12	3	3	12	3	3
subtotal	27	28	30	41	28	30
Inferred data, based on 60-day generation time:						
Blood parameters	33	66	66	19	38	38
Total mitoses per gen.		94	96		66	68
Generations per year		6.1	6.1		6.1	6.1
Total mitoses per year		572	584		401	414

2. Assuming 180-day generation time

	Days	Min. mitoses	Max. mitoses	Days	Min. mitoses	Max. mitoses
Oocyst to salivary gland	10	10	12	22	10	12
Liver	5	15	15	7	15	15
Gametocytes	12	3	3	12	3	3
subtotal	27	28	30	41	28	30
Inferred data, based on 180-day genome time:						
Blood parameters	153	306	306	139	278	278
Total mitoses per gen.		334	336		306	308
Generations per year		2	2		2	2
Total mitoses per year		677	681		621	625

Picking extreme values from **1** and **2** (in red), *total mitoses per year* = 401 to 681

Using data from Claessens *et al*[3]:

Average mutation rate = **3.83 x10^{-10}** per base **per 48 hr cycle**

(equivalent to 1.64 mutations per genome per year *in vitro*)

=>Average mutation rate = **9.57 x10^{-11}** per base **per mitosis**

=>Expected **number of mutations** = (9.57 x10^{-11} x **401**) to (9.57 x 10^{-11} x **681**)

= 3.84 x 10^{-8} to 6.52 x 10^{-8} **per base per year**

= 0.9 – 1.5 **per genome per year** (considering a genome size of 23.3 Mb)

According to Bopp et al[9], excluding parasites grown in presence of drug, the numbers of measured mutations per genome per year were 5.046, 1.682 and 1.682 depending on the isolate. The median value from this study is also nearly identical to that described by Claessens *et al*[3]**.**

In-vivo data

In the *Plasmodium falciparum* IT[10] genome, we observed a region of around 225 – 312 kb, covering the PfCRT locus and an internal *var* gene cluster that is highly conserved in a number of field isolates. Since all of these isolates have the chloroquine resistant genotype, the conserved region is likely to have resulted from chloroquine-selective sweep and could be around 50 years old[11]. However, the presence of a *var* gene on the opposite strand differentiates these isolates from others and may have decreased overall recombination rates in this region.

We called SNPs in this region from 5 isolates produced using PacBio sequencing data from the unpublished Pf3k project, available from: ftp://ftp.sanger.ac.uk/pub/project/pathogens/Plasmodium/falciparum/PF3K/PilotReferenceGenomes/) The detected region that is almost SNP free is shown in Supplementary Fig. 2 and the table below. We observed between 0-10 substitutions per year, with a median of 1.67 mutations per genome per year.

SNPs were called with *mpileup* and *varfilter* from samtools[12], after remapping the reads with BWA[13].

Conserved regions around the PfCRT in five clinical *P. falciparum* isolates.

Assuming that the selection occurred ~50 years ago, we obtain the reported estimate of mutations per year and genome.

Isolate	Country	Location of valley on PfIT_07	Size of valley	mpileup SNPs	Manually inspected SNPs	estimated mutation rate / year / genome
SenT128.08	Senegal	383941..609688	225	0	0	0
PA0085-C	The Gambia (1)	378352..699841	312	1	1	1.57
PM0138-C	Mali	372520..664849	292	1	1	1.68
PA0012-C	The Gambia (2)	378563..660686	283	6	6	10.38
PD0469-C	Thailand	228905..510000	281	4	2	3.48

Samples are from the Pf3K pilot project. For further details see ftp://ngs.sanger.ac.uk/production/pf3k/release_5/pf3k_release_5_metadata_20170804.xlsx

In conclusion, we assume that *P. falciparum* accumulates on average 0.9-1.5 SNPs per genome per year. We assume that this value is also valid for the other *Laverania* species.

Coalescent models

To estimate key population genetic parameters: effective population size (N_e), dates of divergence (*D*) and number of migrants per generation from source population to target population (*M*), we used the Generalised Phylogenetic Coalescent Sampler (G-PhoCS)[1]. As input, we used 1750 alignments from the Lav15sp dataset. We ran two models: (1) split between *P. falciparum* and *P. praefalciparum*, and (2) the entire *Laverania* tree. We incorporated phylogenetic information and modelled bi-directional migration between all extant and ancestral nodes. The MCMC chains were run for a minimum of 10 million iterations, with 20 chains run in parallel. The chains were merged and manually checked for convergence (Tracer version 1.5). We estimated $N_e = \vartheta / 2\mu$, $D = g*\tau / \mu$ and $M = m_{st} * \tau$, where ϑ, τ and m_{st} *(migration source to target)* are model parameters, μ is the mutation rate per base pair per generation (ranges from $6.952\text{x}10^{-9}$ to $1.158\text{x}10^{-9}$ per base-pair per generation, equivalent to 0.9 - 1.5 mutations per year per genome) and *g* is the generation time of 0.18 to 0.5, as described above. The *M* parameter is estimated as the total migration rate, approximately indicating the probability that a given lineage in the source population will migrate into the target population[14]. This migration can be seen in some cases (see Supplementary Table 3) especially from *P. praefalciparum* into *P. falciparum*.

We applied the algorithms to three types of alignments (see Supplementary Table 3): (1) genic regions and (2) intergenic regions with and without assumed 500-bp untranslated regions. These alignments appeared robust for the *P. reichenowi*, *P. praefalciparum* and *P. falciparum* comparison as well as for *P. adleri* and *P. gaboni.* However, alignment of more distantly related species was not possible due to a high number of insertions and deletions and the low GC content. We performed the dating on genic alignments for all possible species for which we had more than 2 samples (thus excluding *P. blacklocki*). For the estimates used in Fig. 1 and Supplementary Table 3, some of the estimates of population genetic parameters were approximated where we were unable to generate intergenic alignments.

Multiple Sequentially Markovian Coalescent

To estimate changes in effective population size (*Ne*) over time in *P. falciparum* (PfGA01 & PfIT, from Pf3K dataset), *P. praefalciparum* (PprfG02 & PprfG01) and the gene-flow between them, we ran the multiple sequentially Markovian coalescent (MSMC) on segregating sites from all chromosomes[2]. Genome-wide SNPs were generated by firstly mapping raw reads from each sample against the Pf3D7 reference, then piping BAM files through mpileup v. 0.1.9 (parameters -q 20 -Q 20 -C 50) into bcftools call v. 1.1 (see MSMC documentation for more details). Retaining only homozygous SNPs, each *Plasmodium* chromosome was considered a single phased haplotype. MSMC was run for 20 iterations with a fixed recombination rate. Effective population size was calculated as $(1/\lambda)/2\mu$, scaled time was converted into years as $(scaled\ time\ /\ \mu) * g$. The parameters λ and scaled time are derived from the model. Values for parameters μ *and* g are described above. The error around our estimates was estimated by bootstrapping 50 replicates by randomly resampling from the segregating sites used as input.

Estimation of population size

Effective population size was estimated from 10,000 years before present (BP) until 500 years BP as bootstrapping demonstrated that the model loses resolution for more recent periods than 500 years BP. The effective population size of *P. falciparum* drops from at least 11,000 years BP and steadily declines to reach its lowest value around 6,000-4,000 years BP ($N_e \sim 3000$), before the population size begins to expand thereafter until 500 years BP (Fig. 2b). While others have speculated on the census population size of *P. falciparum* at this time[15] there is no straightforward way to relate N_e to census population size (N) due to complexities in the life-cycle of *P. falciparum* that causes the population to deviate from certain assumptions of the Wright-Fisher model[16]. Nevertheless, generally the census number of parasites is much higher than N_e[17]. The bottleneck is unique to *P. falciparum*; although there is a large degree of error in the bootstrapping around N_e estimates for *P. praefalciparum*, the gorilla-infective species does not appear to go through a bottleneck during this period. We replicated the analysis with different *P. falciparum* genomes (PfDd2 & PfHB3, from the Pf3K dataset), which produced near-identical results.

Based on evidence from selection in the human genome, the origin of human malaria has been estimated as ~40,000–60,000 years BP and a population expansion associated with the origins of agriculture is assumed to have taken place ~4,000–6,000 years BP[18]. This scenario is confirmed by our modelling of the speciation event between *P. falciparum* and *P. praefalciparum* with G-PhoCS estimates of the timing of speciation ranking from 40,000–60,000 years BP [30,000–70,000 (95 % CI)] and the MSMC estimates of N_e through time showing a rise from 4,000-6,000 years BP onwards.

Dating of *eba-175* dimorphism

To adapt the dating of the *eba-175* dimorphism[19], the following calculation was performed. Previous authors used 6 million years BP as the time when *P. reichenowi* split from the ancestor of *P. falciparum* and *P. praefalciparum* and then dated the *eba-175* split to 0.13–0.14 MYA. As those numbers can be scaled linearly, we used time of 0.13 - 0.23 MYA for the *P. reichenowi* split, which puts the data of the *eba-175* split to around 3,000-5,000 thousand years ago. This agrees with our observation that the dimorphism of *eba-175* occurred in *P. falciparum*, not *P. praefalciparum* (Supplementary Fig. 3b), concluding that the dimorphism occurred during the expansion of the *P. falciparum* and its host.

Supplementary Note 2: Evolution of core genes

Within-species polymorphism

The nucleotide diversity per CDS (π), the average number of nucleotide differences per site between two sequences, was calculated for each species (Supplementary Fig. 1) and their means compared using non-parametric Wilcoxon rank sum tests. Differences in the observed nucleotide diversity may reflect variation in prevalence and different demographic histories of the great ape parasites. The *P. falciparum* nucleotide diversity (computed from 5 worldwide isolates) was significantly lower than the nucleotide diversity observed in any other great ape species (calculated from 2 to 5 genotypes collected in the same localization from Gabon) ($p < 0.0001$). All Wilcoxon rank sum test results comparing nucleotide diversity between the parasites of great apes were significant ($W = 48193000$, $p < 0.0001$; Supplementary Fig. 1). The nucleotide diversity observed in gorilla-infecting species was higher than the diversity observed in the chimpanzee-infecting species ($W_{P.praefal.\text{-}P.\ adleri} = 4963900$, $p < 0.0001$). Among the gorilla-infecting species *P. praefalciparum* presented higher diversity than *P. adleri* ($W_{P.\ adleri.\text{-}P.\ praefal.} = 9540900$, $p < 0.0001$), due to a higher number of genes with relatively high values of nucleotide diversity. When considering only genes with a nucleotide diversity ≤ 0.02, the diversity was higher in *P. adleri* ($W_{P.\ adleri.\text{-}P.\ praefal.} = 9540900$, $p < 0.0001$). Regarding chimpanzee-infecting species, the diversity was significantly higher in *P. gaboni* ($W_{P.\ reichenowi\text{-}P.\ gaboni} = 5719500$, $p < 0.0001$) and lower in *P. billcollinsi* ($W_{P.reichenowi\text{-}P.\ billcollinsi} = 8011900$, $p < 0.0001$). The lowest diversity was observed in the least prevalent species, *P. billcollinsi* that infects chimpanzees.

Interspecific gene transfer

Most of the CDS topologies (4,319 out of 4,350, 99.6%, "Lav7sp" dataset) did not significantly differ from the *Laverania* species tree. For the remaining CDS (n=31, including 4 genes of chromosome 4, Supplementary Fig. 5), we specifically looked at their topology and identified those with possible events of gene transfer between species parasitizing the same host species. We detected a clustering of divergent species infecting the same host for eleven CDS, but none of them included all the species infecting the same host, *i.e. P. adleri*, *P. blacklocki* and *P. praefalciparum* or *P. gaboni*, *P. billcollinsi* and *P. reichenowi*. Four of them, localized in the same region of the chromosome 4, shared the same topology, with *P. praefalciparum* and *P. falciparum* grouping together with *P. adleri*, and corresponded to the previously reported introgressed genomic island (topology B in Supplementary Fig. 4; see main text). In the other cases, the chimpanzee-infecting species *P. billcollinsi* was closer to clade A species (four genes; topology C in Supplementary Fig. 5) or clustered together with *P.*

reichenowi (3 genes; topology D in Supplementary Fig. 5). All these signals remained when considering all sequenced genomes (dataset Lav15st) and concerned in some instances the intergenic region too (see Supplementary Fig. 5, the table below). Beyond these cases, most often, deviations of gene tree topologies from the species tree involved a clustering of *P. billcollinsi* and/or *P. blacklocki* closer to *P. adleri* and/or *P. gaboni* compared to the species phylogeny (not shown), or concerned alignments without enough resolution.

Genome-wide test of convergent evolution

We searched for an excess of convergent substitutions in specific branch-pairs by analyzing the correlation between the number of convergent and divergent substitutions between all the branch-pairs in a phylogeny, and looking for outlier branch-pairs that had high positive residuals, indicating an excess of convergent substitutions relative to the number of divergent substitutions[20]. Both for the divergent and convergent substitutions and for all pairwise comparisons, Pearson's correlation coefficients between the number of substitutions estimated under distinct evolutionary models were always higher than 0.99. We therefore only report the results obtained under the LG model of amino-acid substitutions. At a chromosome scale, we did not detect an excess of convergence between parasite species infecting the gorillas or between the parasites infecting the chimpanzees. However, we detected an excess of convergent substitutions relative to divergent substitutions, in three branch-pairs involving *P. blacklocki* but with no association with the host species.

Supplementary Note 3: Gene family analyses

Differences in gene families

The *P. reichenowi, P. gaboni* and *P. adleri* reference genomes are from single *Laverania* infections where a single isolate predominated (see Supplementary Table 1). However, the *P. praefalciparum* sample contained two distinct genotypes of *P. praefalciparum*. For the core region, a single haploid assembly could be resolved into the two genotypes. For more variable regions of the genome, like the subtelomeres, the genotypes could not be completely resolved and the numbers reported for the *rif*, *stevor* and *var* genes therefore contain contributions from both haplotypes. For *P. billcollinsi* and especially for *P. blacklocki*, we could not estimate the extent to which the subtelomeres assembled. Although gene families, like CLAG and the *var* genes from internal clusters, did assemble, the numbers of variable genes families are likely to be underestimated due to amplification biases introduced by the sWGA approach for *P. blacklocki* and the fact that *P. billcollinsi* is obtained from a co-infection with *P. gaboni*.

To estimate the number of genes we used (a) a regular expression to count the genes based on functional annotation and (b) matches to Pfam domains (E-value < 1e-6). To each gene/domain we associated counts and standard deviations (Supplementary Tables 6a, b). Differentially distributed gene families are reported in Fig. 3. For several genes, we performed phylogenetic analyses (Supplementary Fig. 7) to better understand their evolution. This was done by aligning the genes of a specific group with Muscle[21] using default parameters. In Seaview[22], we ran GLOCKS[23] with permissive settings and PhyML[24] (default settings for amino acids) to construct trees. The obtained trees were analysed in Figtree[25].

To perform the alignments of *msp* and *eba-175* dimorphic alleles, the same method was used but the numbers of sequences were reduced by subsampling to visualize dimorphisms.

Generation of similarity matrices

Where sequences were too divergent to perform tree based analyses, we implemented a visualization method based on similarity scores. First, amino acid sequences were compared with a BLASTp (e-value < 1e-6 and low complexity filter set to false). A similarity matrix based on the score or the global identity was built (the alignment length was normalized by the mean sequence length). Using the similarity matrix, the aligned sequences were clustered using the ward.D2 algorithm in the heatmap.2 module of gplots in R[26]. To each gene, we associated their species and in some case their functional annotation through further heatmaps. We used this approach to analyse domains of *var* genes (see below, Supplementary Fig. 10).

The Rifin and Stevor proteins

To build a BLAST-based network, all Pir proteins were compared with an all-against-all BLASTp (parameter: -e 1e-6 –F F). We clustered the Pir proteins using Gephi[27] and tribeMCL[28] into groups, used in Fig. 3. For the Stevor proteins, we built a phylogenetic tree, using RAxML, with the PROTGAMMAIGT model and 100 bootstraps.

Meme-Motif analysis for Stevor proteins

To predict motifs in this family, we used MEME[29] version 4.9.1. We searched for 96 motifs of 8-15 amino acids using all of the Stevor proteins encoded by the seven reference genomes. Proteins with less than 5 hits were excluded. The output was parsed with a PERL script into a matrix and visualized in R[26], using the heatmap.2 function and the ward2 clustering (Supplementary Fig. 8).

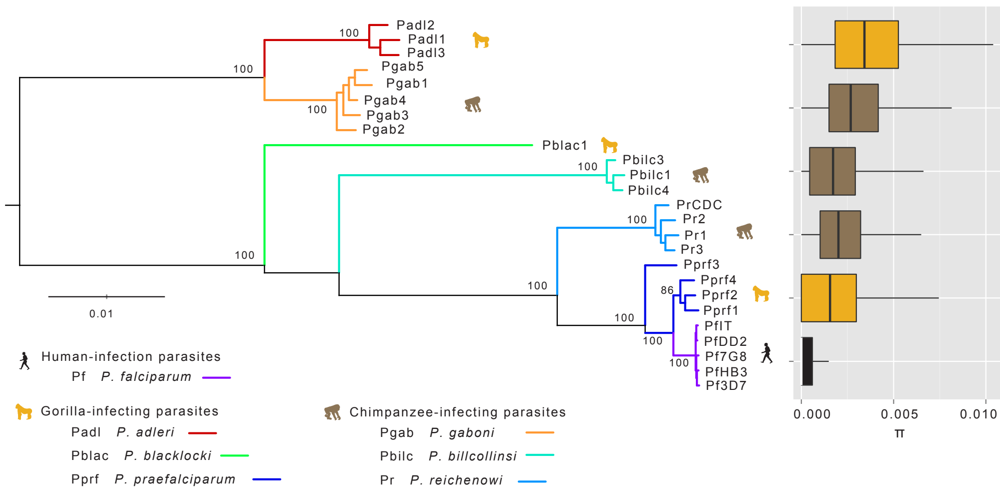
var gene analysis

To analyse full-length *var* genes in the *Laverania* we excluded genes smaller than 2.5kb and called domains in the genes. The following domains were identified from their conceptual translations: ATS (Acidic Terminal Sequence), NTS (N-Terminal Sequence), DBL (duffy binding like), CIDR (cysteine-rich interdomain region), pam (placenta associated malaria) and the duffy-binding like domain as defined by Pfam (present in invasion related proteins). To call domains, the program hmmscan[30] was used with the HMMer models from the VARdom server using the following parameters: --domT 50 -E 1e-6 to attribute domains to *var* genes. As the domains are similar to each other, we generated a PERL program that ascribed domains based on best scores (at least 80% of the length of the HMMer domains). The regions of *var* genes encoding domains could overlap by up to 20 bp. In some cases, rather than finding one of the known domains (DBL, CIDR, ATS or NTS) the Pfam-defined duffy binding-like domain was found. If this happened, we named that domain Duffy, rather than Duffy Binding-Like. Regions (≥ 300aa) in the *var* genes not covered by known domains were also extracted and first called “Unclassified”. From those “Unclassified” domains a novel domain was found that we termed CIDRn because of the similarity to existing CIDR domains (Supplementary Fig. 9). To better understand the structure of the domains, particularly Duffy, we used a similarity matrix (Supplementary Fig. 10c).

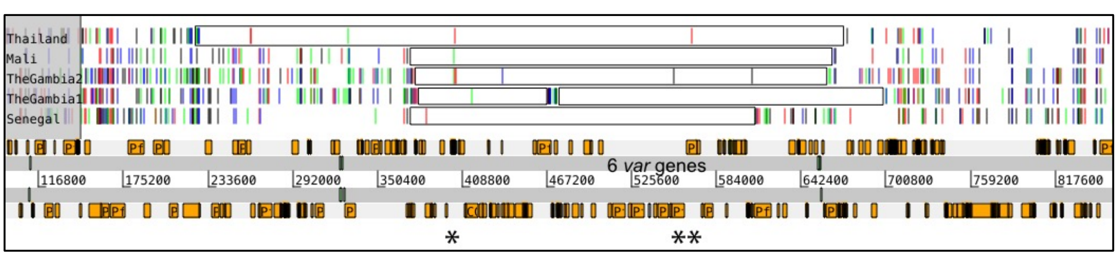
It can be seen that some domains like CIDRα or ATS form defined groups, with little similarity to others. Other domains like DBLα with DBLβ share sequence similarity. The distribution of DBLε, DBLpam2/3 and the unclassified Duffy domain is noteworthy (dotted black lines, top of Supplementary Fig. 10c). These domains seem to be most common in *P. praefalciparum*, *P. adleri* and *P. gaboni*. They have less similarity to other domains. Rather than representing a new domain (like a DBLx[31]), we think that those domains might be more ancient.

We also classified the DBLx domain proposed by Larremore *et al*[31]. Their sequences that started with the amino acids specific for the DBLx domain (start NI or DF, end CPQNLDFDRRDQFLR) were compared to our domain dataset. Domains containing those sequences were labelled as DBLx in our set. Next, we generated a similarity matrix with those DBLx, DBLε and Duffy (Supplementary Fig. 12) The DBLx labeled sequences are clustered within the DBLε group. Therefore, we think that the DBLx is not a new domain, but rather part of the diverse DBLε group.

Supplementary Figures

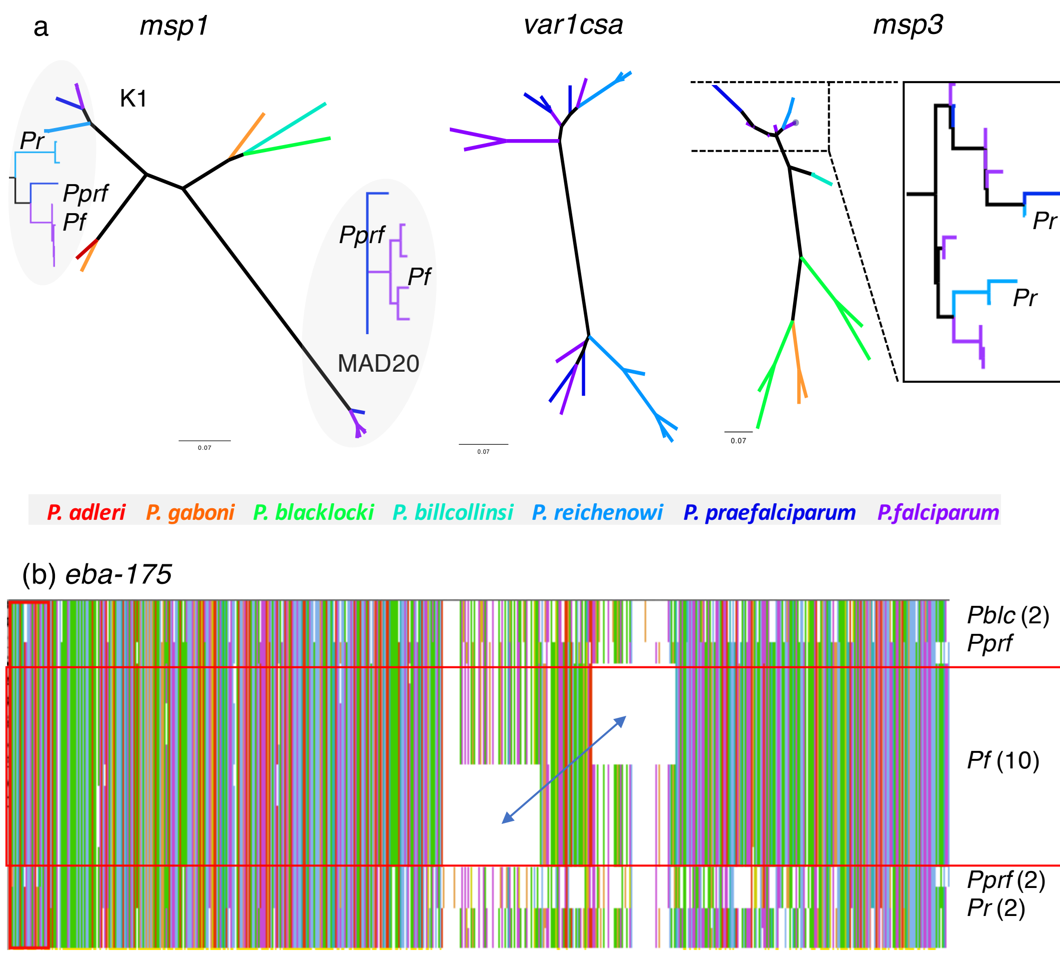


Supplementary Fig. 1. Maximum Likelihood tree and nucleotide diversity of *Laverania* isolates. The tree was obtained using the sequences of 424 genes ("Lav25st" set of orthologues). The box plots show the nucleotide diversity per CDS (π) for each species. Each boxplot (Tukey's box plot: median, 25th & 75th percentiles and the whiskers extend to the farthest points that are within 1.5 times the interquartile range) is based 3,808 comparisons ("Lav15st" data set).

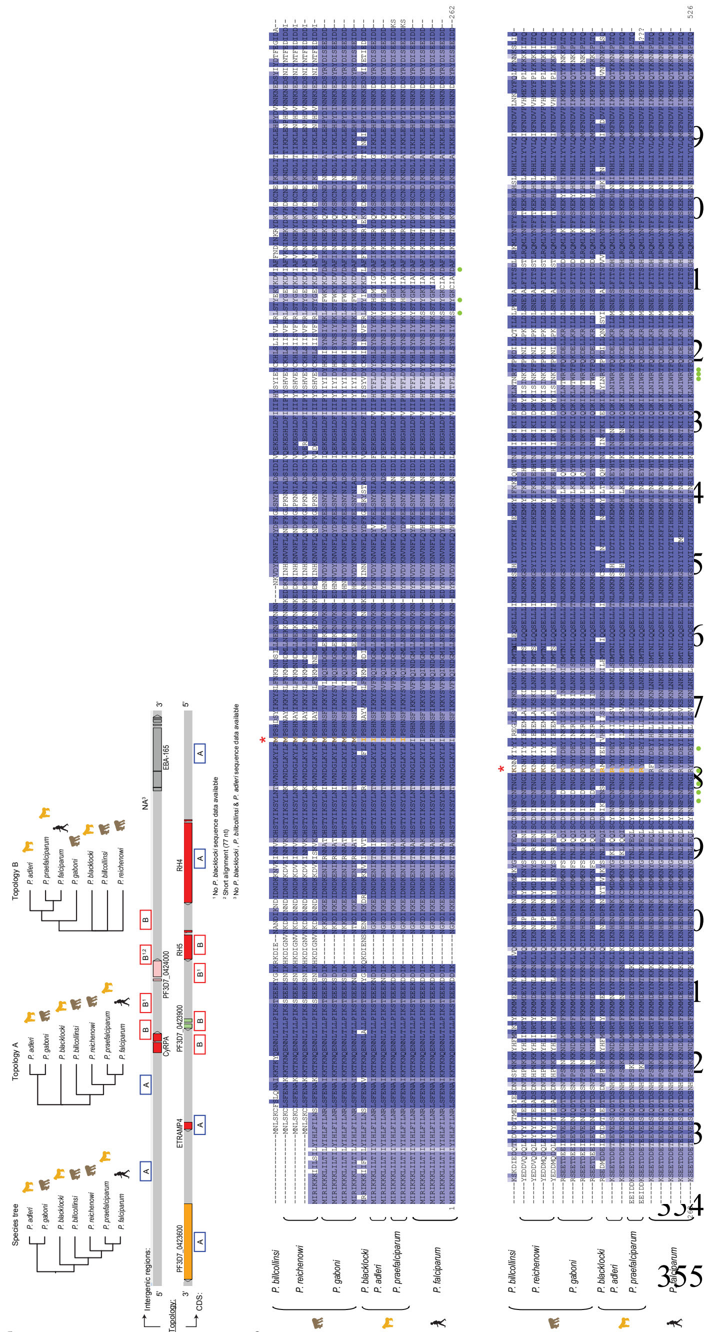


Supplementary Fig. 2. Estimation of *in-vivo* mutation from field isolates. An Artemis view of the 600 kb conserved region of five clinical *P. falciparum* isolates (Thailand, Mali, The Gambia 1 and 2 and Senegal), around the *Pf*CRT(*) locus.

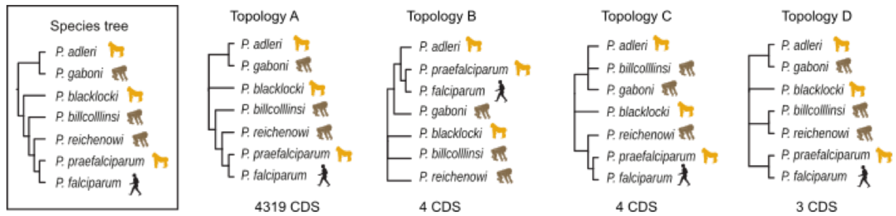
Orange boxes represent the annotated genes on both DNA strands along chromosome 7. For each isolate, variation of nucleotide sequences (SNP) compared to the *P. falciparum* 3D7 reference genome are indicated by coloured bars. Large, nearly SNP-free, regions (black boxes) of around 200 kb are found. One *var* gene (**) in the internal cluster is on the opposite strand.



Supplementary Fig. 3. Dimorphisms in the *Laverania*. (**a**) Examples of ancient dimorphisms based on maximum likelihood phylogenetic trees. Dimorphism in *msp1* arose in the *P. falciparum–P. praefalciparum* ancestor, after the divergence of *P. reichenowi* and dimorphism in *var1csa* evolved in the *P. reichenowi–P. praefalciparum–P. falciparum* ancestor after the divergence of *P. billcollinsi*. There is also evidence of a bi-allelic distribution of *msp3* in *P. falciparum, P. praefalciparum* and *P. reichenowi*. (**b**) Dimorphism in *eba-175* is more recent. The alignment shows two mutually exclusive indels (arrow) in the *P. falciparum* sequences, not present in other *Laverania* species. The colours represent different nucleotides. For the *P. falciparum* sequences, we used full sequences from the following Pf3K isolates: PfML01, PfSD01, PfDd2, Pf7G8, PfHB3, PfSN01, PfIT, PfCD01, PfGB4 and PfGN01). *Pf*, *P. falciparum; Pprf, P. praefalciparum; Pr, P. reichenowi*; and *Pbilc, P. billcollinsi*

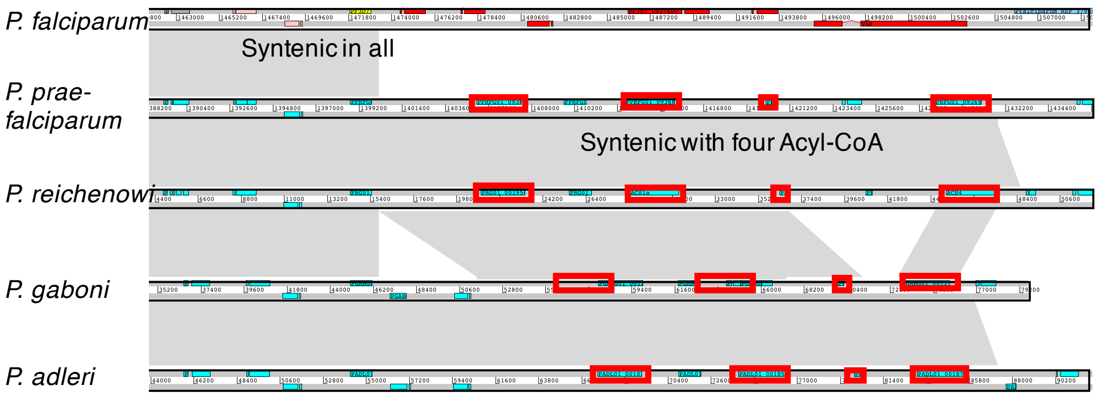


Supplementary Fig. 4. Interspecific gene transfer and convergent evolution at the right end of the chromosome 4. (**a**) Support for interspecific gene transfer between the gorilla-infecting species *P. adleri* and the common ancestor of *P. praefalciparum* and *P. falciparum*. The topologies observed in the coding and intergenic regions of the end of chromosome 4 and beyond are given. (**b**) Convergent evolution in the *rh5* gene. Amino acid alignment of the *rh5* region that carries the significant fixed difference between parasites infecting the chimpanzees and those infecting gorillas (red stars). Green circles indicate positions that are known to be involved in the interaction with the human receptor Basigin[32].

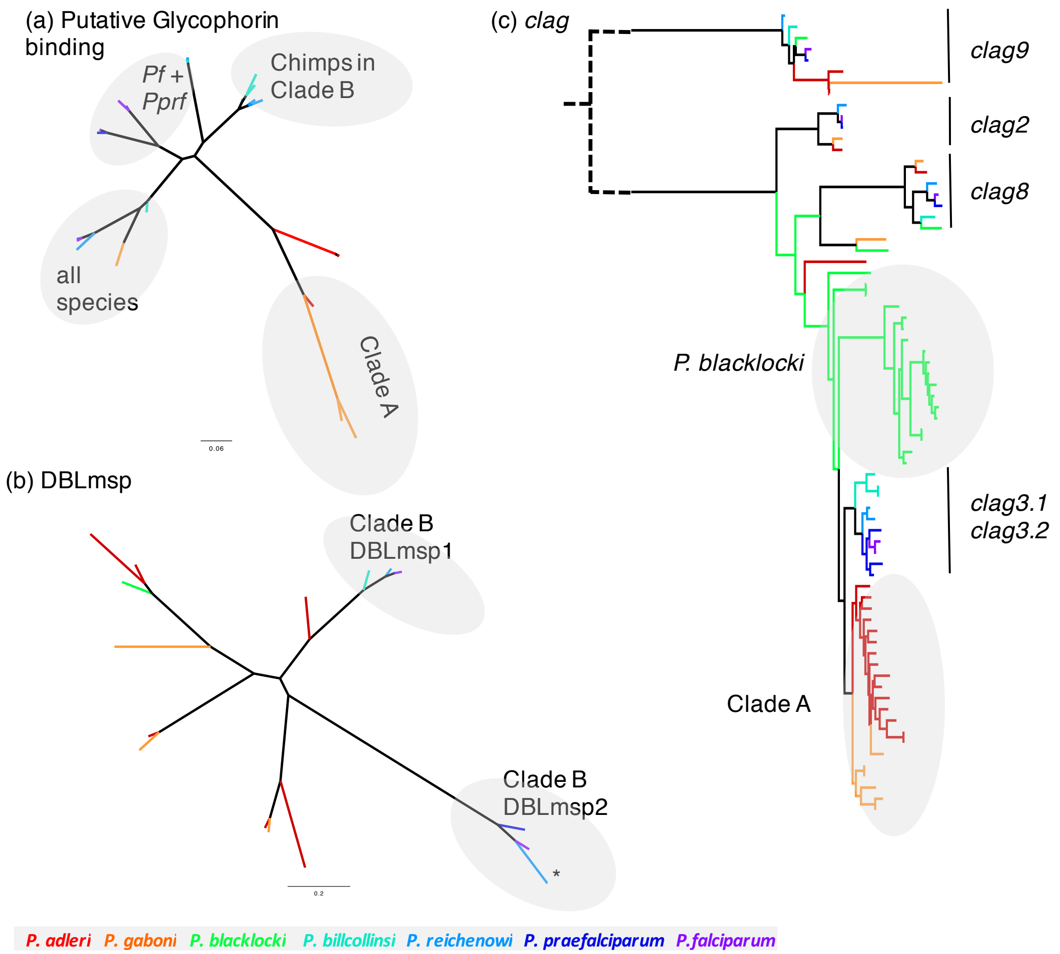


gene ID	function	gene name	topology	newick tree	all strains		intergenic regions		note
					nt	aa	5'	3'	
PF3D7_0423800	cysteine-rich protective antigen RH5-Ripr membrane anchoring protein	CyRPA	B	(((PADL01,(PPRFG01,PF3D7)),PGABG01),(PBILCG01,PRG01),PBLACG01);	yes	yes	no	yes	(Pprf, Pf) clusters together with Padl
PF3D7_0423900	probable protein, unknown function		B	(PBILCG01,((PGABG01,((PF3D7,PPRFG01),PADL01)),PRG01),PBLACG01);	yes	yes	yes	yes	(Pprf, Pf) clusters together with Padl
PF3D7_0424000	Plasmodium exported protein (PHISTc), unknown		B	((PGABG01,(PADL01,(PF3D7,PPRFG01))),PBILCG01,PRG01);	yes	yes	yes	yes	No Pblac sequence (Pprf, Pf) clusters together with Padl
PF3D7_0424100	reticulocyte binding protein homologue 5	RH5	B	((((PPRFG01,PF3D7),PADL01),PGABG01),(PRG01,PBILCG01),PBLACG01);	yes	yes	yes	yes	(Pprf, Pf) clusters together with Padl
PF3D7_0524000	karyopherin beta	KASbeta	C	(PPRFG01,(PRG01,(PBLACG01,((PBILCG01,PGABG01),PADL01))),PF3D7);	yes PgabG02 inbetween (Padl,PgabG01,Pbilc) & the others	Pbilc closer to cladeA sp. but outside (Padl,Pgab)	no	NA no Pbilc sequence	Pbilc clusters with Pgab
PF3D7_0613200	conserved Plasmodium protein, unknown function		C	((PBLACG01,((PF3D7,PPRFG01),PRG01)),(PGABG01,PBILCG01),PADL01);	yes	NA	yes	yes	Pbilc closer to clade A
PF3D7_1327500	conserved Plasmodium protein, unknown function		C	((PGABG01,PADL01),(PBLACG01,(PRG01,(PF3D7,PPRFG01))),PBILCG01);	yes	yes	yes	no	Pbilc closer to clade A
PF3D7_1328000	conserved Plasmodium protein, unknown function		C	(((PBLACG01,(PBILCG01,(PADL01,PGABG01))),PRG01),PPRFG01,PF3D7);	yes	yes	no	NA	Pbilc closer to clade A
PF3D7_0102400	lysophospholipase, putative, pseudogene		D	(PADL01,((PF3D7,PPRFG01),((PBILCG01,PRG01),PBLACG01)),PGABG01);	yes	yes	yes	no	Pbilc & Pr cluster together
PF3D7_0102500	erythrocyte binding antigen-181	EBA181	D	((((PADL01,PGABG01),PBLACG01),(PBILCG01,PRG01)),PF3D7,PPRFG01);	yes	yes	no	no	Pbilc & Pr cluster together
PF3D7_0902600	serine/threonine protein kinase, FIKK family	FIKK9.7	D	((((PPRFG01,PF3D7),(PRG01,PBILCG01)),PBLACG01),PADL01,PGABG01);	yes	yes	yes	yes	Pbilc & Pr cluster together

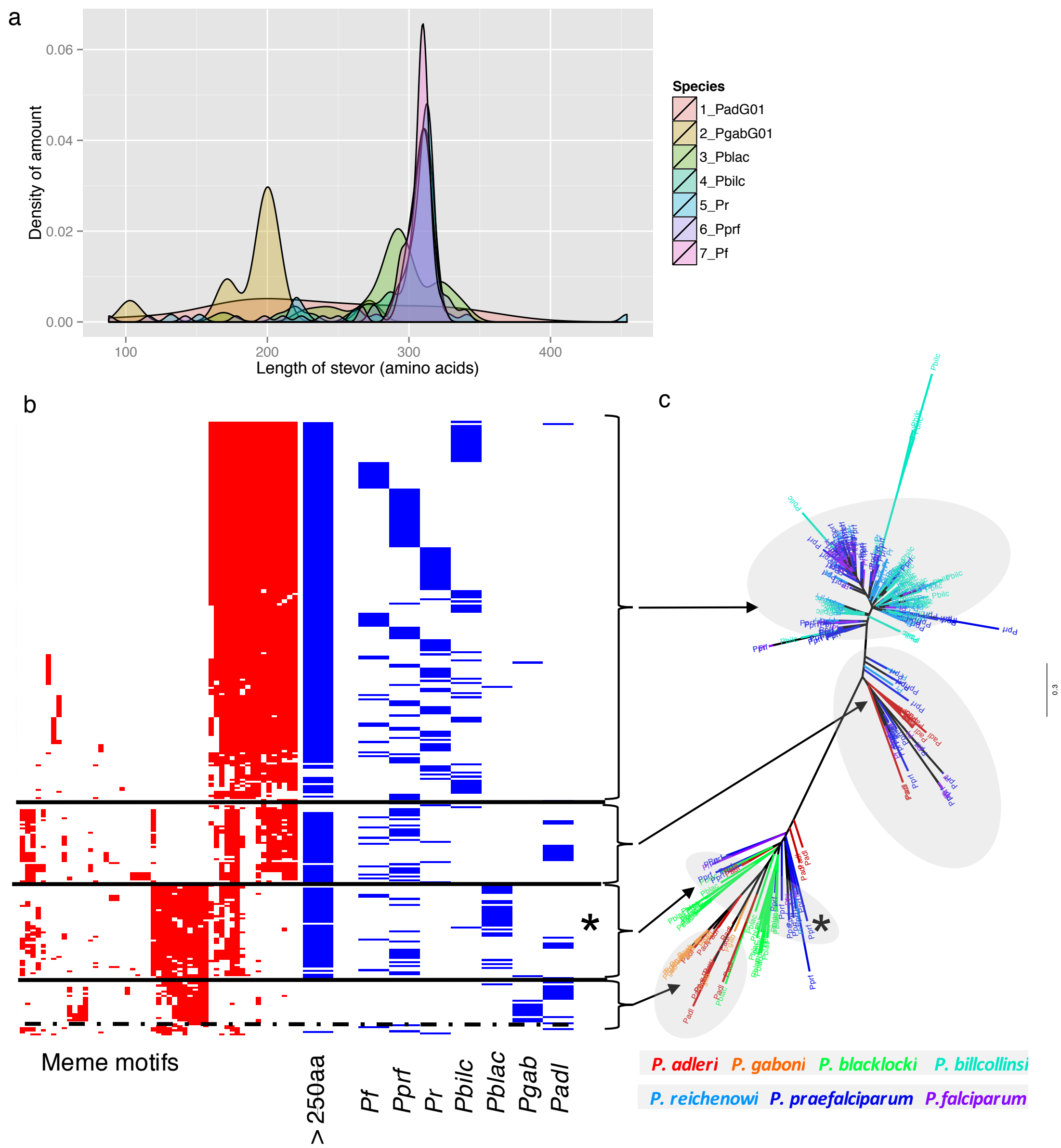
Supplementary Fig. 5. Tree topology tests. The number of protein coding sequences (CDS) producing each of the major tree topologies (observed for >1 CDS) is shown. The table summarizes the results for the 11 CDS with signals of gene flow between species infecting the same host (*i.e.* other signals are not considered here). The table also shows whether a given signal was still observed when all strains were considered, using the nucleotide or amino acid sequences and whether the signal was observed in the intergenic regions down- and up-stream of the respective genes.



Supplementary Fig. 6. Acyl-CoA Synthetase expansion on Chromosome 9. ACT view of five genomes, at the right-hand side of chromosome 9. The grey areas indicate co-linearity. *P. falciparum* has lost this region with four Acyl-coA synthetase genes, as this locus is conserved in the other species.



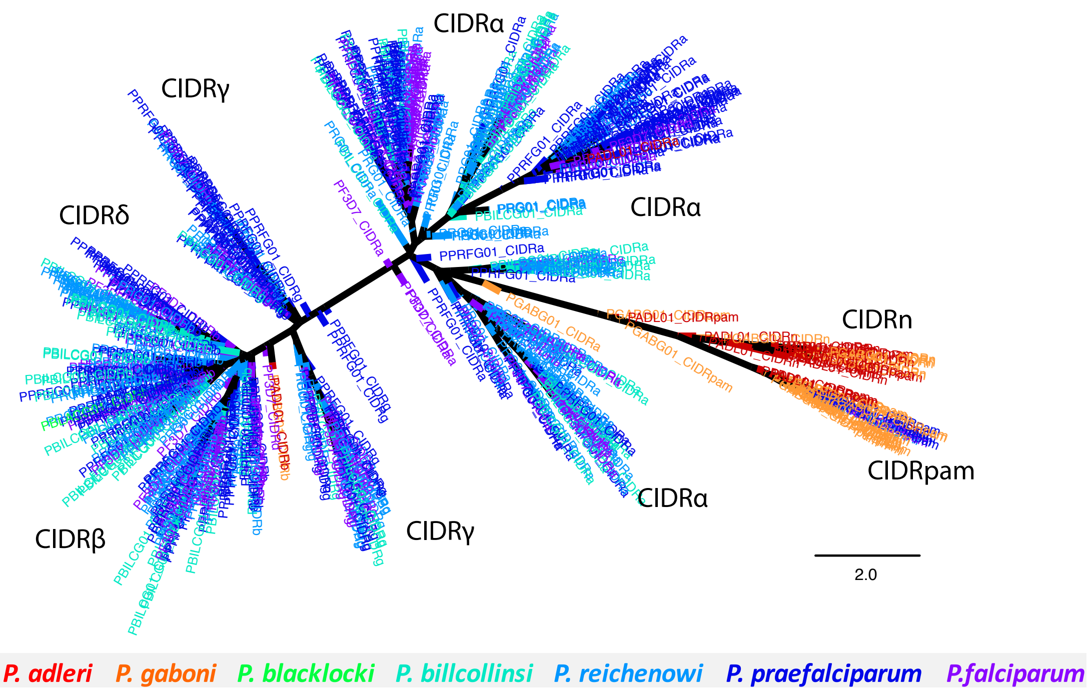
Supplementary Fig. 7. Phylogenetic analysis of multigene families. Example of three families that show differences within the *Laverania*. (**a**) The putative glycophorin binding proteins form four distinct groups. One group contains sequences from all species. The remaining groups are clade, host or species sub-group specific. (**b**) Differences in the DBLmsp that are expanded in Clade A. The DBLmsp2 is a pseudogene (*) in *P. reichenowi*. (**c**) Expansion of *clag* genes in Clade A. The distance between the CLAG9 clade and its nearest neighbour has been compressed to aid visualisation (dotted lines).



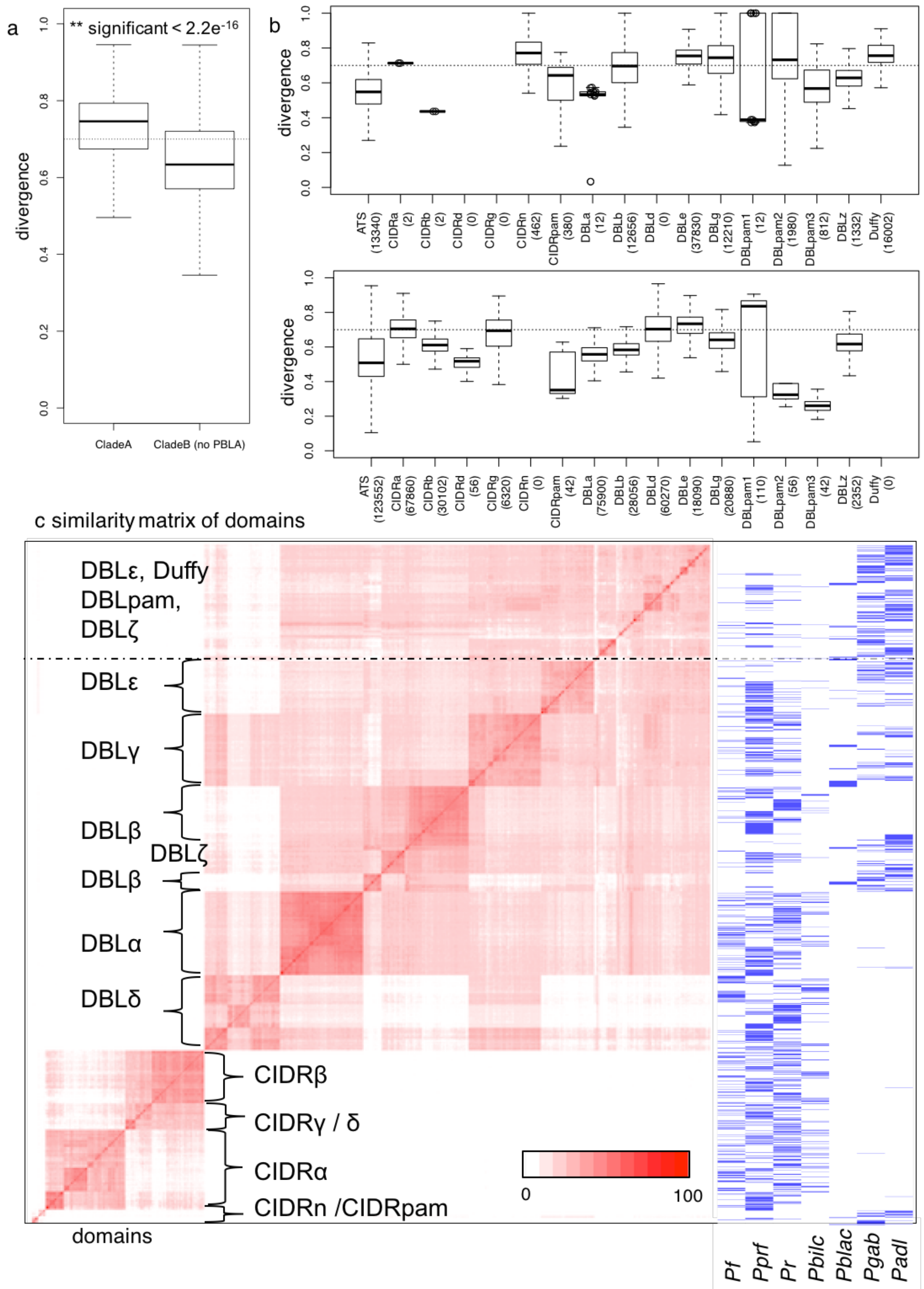
396

397 **Supplementary Fig. 8. Analysis of Stevor proteins.** (**a**) Length of Stevor proteins
398 for the seven *Laverania* genomes. (**b**) Occurrence matrix of meme motifs generated
399 for Stevor proteins (Supplementary note 3). Columns represent the different meme
400 motifs, rows represent all the 301 Stevor proteins. To classify each gene, a binary
401 barcode (blue) is used to indicate whether it encodes likely full length protein
402 (>250aa) and to indicate the species in which it is found. The matrix was clustered
403 with the ward2 algorithm. Note that one cluster (*) has no full length Stevor proteins
404 in chimpanzee parasites. *Pf, P. falciparum; Pprf, P. praefalciparum; Pr, P.*

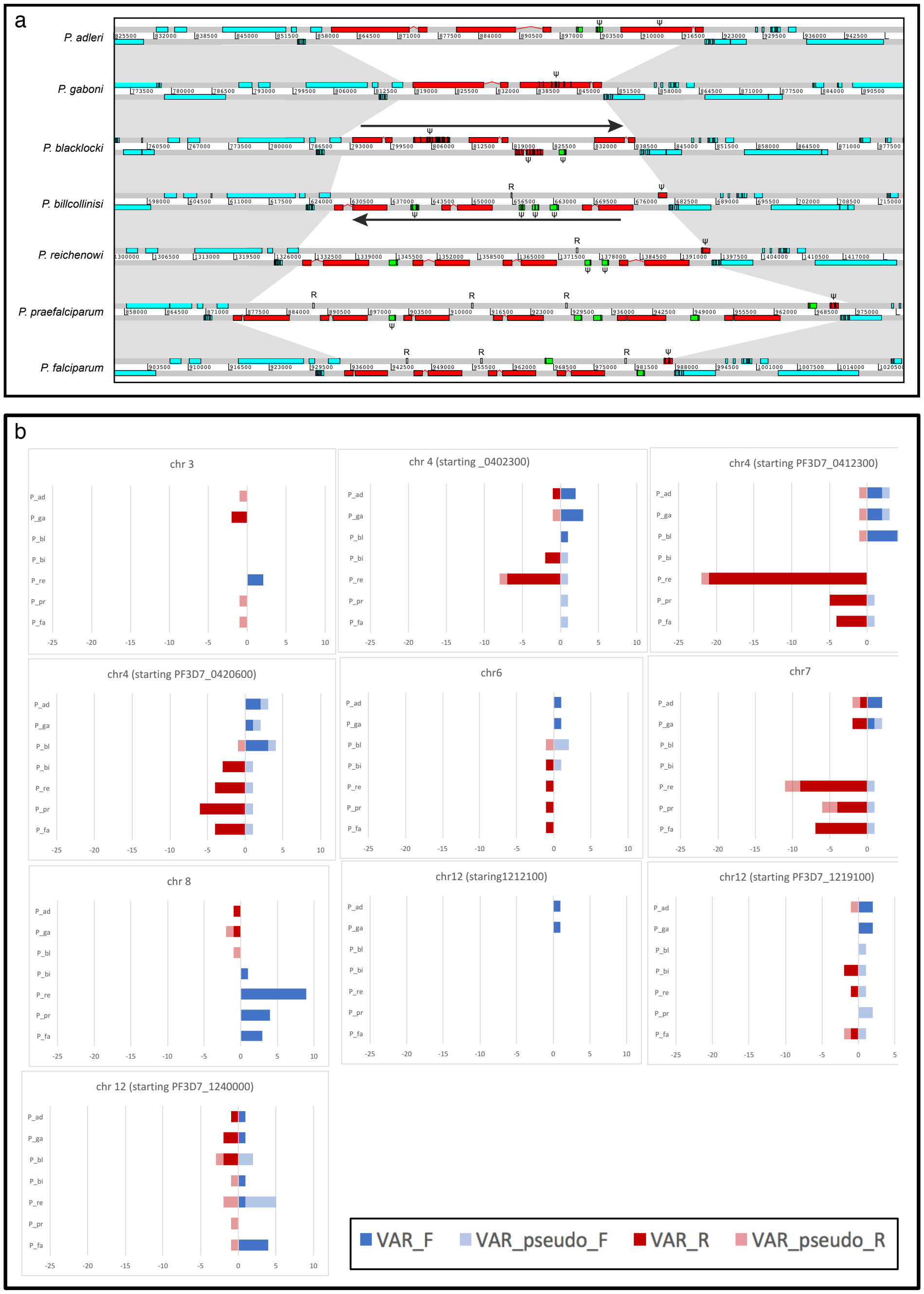
reichenowi; Pbilc, P. billcollinsi; Pblac, P. blacklocki; Pgab, P. gaboni; and *Padl, P. adleri*. (**c**) Maximum likelihood tree of the same data. Bootstrap values of 100 were obtained for all branches.



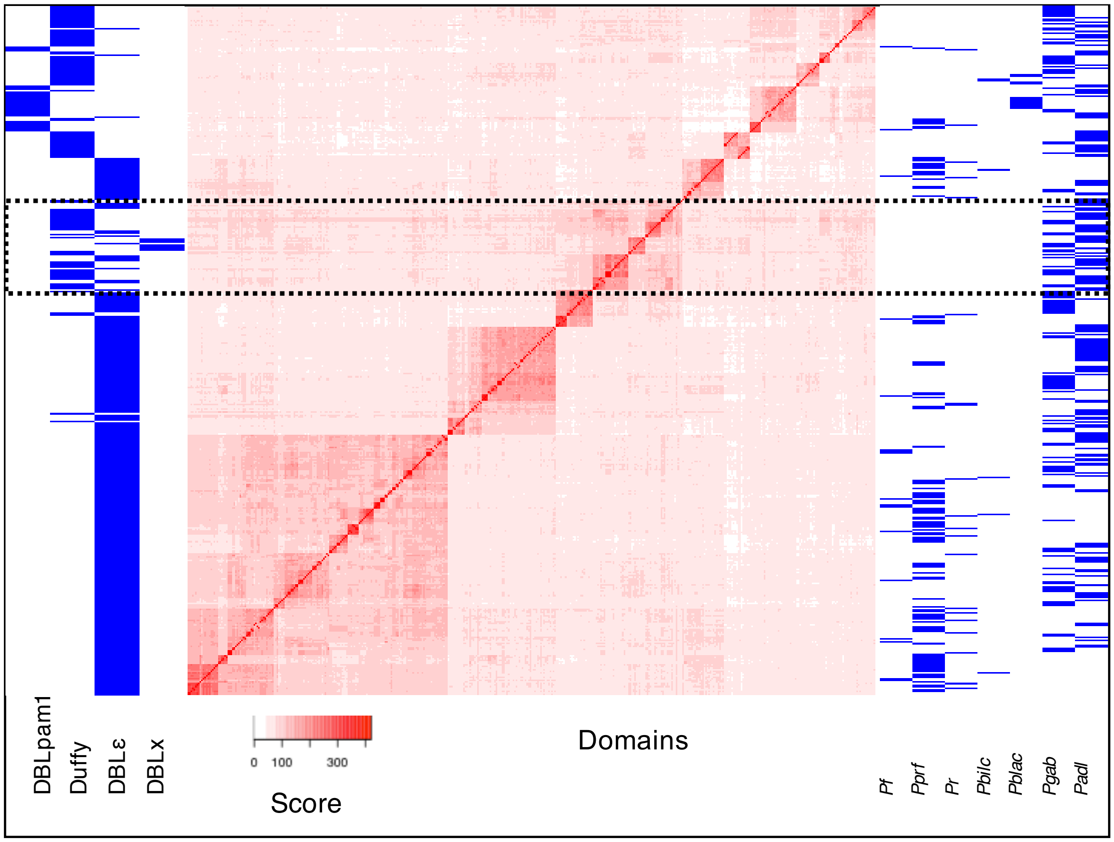
Supplementary Fig. 9. Phylogenetic position of the new CIDR domain (CIDRn) specific to Clade A parasites relative to other CIDR domains. Phylogenetic tree was obtained with RAxML using the PROTGAMMAIGTR models. Bootstrap values of 100 were obtained on all branches.



Supplementary Fig. 10. Diversity of *var* genes domains. (a) Relative domain similarity between Clade A and Clade B (excluding *P. blacklocki*) based on the average across all domains except ATS. The difference observed between Clade A and Clade B is statistically significant (t-test, two sided). Boxplots are based on 83,692 and 310,136 comparisons. (b) Relative similarity across all *var* domain types in Clade A (top) and Clade B (bottom, excluding *P. blacklocki*). The number of predictions for each domain type are shown in parentheses. (c) Annotated similarity matrix between all *var* domains (> 220aa) of the *Laverania* as defined in Fig. 5, including their species and cluster attributions. The similarity matrix shows the score of the BLASTp between the domains, clustered with the ward2 algorithm in R. Each row and column represents one domain and shows its similarity to the other 2,467 domains (and itself). The occurrence of each domain/row across the species is indicated by the blue bars on the right. Although domains above the dotted line are classified differently, they cluster together. Pf, *P. falciparum*; Pprf, *P. praefalciparum*; Pr, *P. reichenowi*; Pbilc, *P. billcollinsi*; Pblac, *P. blacklocki*; Pgab, *P. gaboni*; and Padl, *P. adleri*. All boxplots are Tukey's box plots, showing the median, 25th & 75th percentiles and the whiskers and the whiskers extend to the farthest points that are within 1.5 times the interquartile range.



Supplementary Fig. 11. Composition, structure and evolution of *var* genes within *Laverania*. (a) Screenshot from ACT showing the 2nd internal cluster of *var* genes on chromosome 4 in the seven *Laverania* species. In Clade A and *P. blacklocki*, the orientation of the *var* genes is different compared to that of the other species. The GC-rich RUF elements[33] (RNA of Unknown function), highlighted with an R, occur less frequently in Clade A genomes. The size of the *var* genes between the species is different. *var* genes are in red, *pir* genes in green. (b) Bar plot of the number and orientation of *va*r genes or pseudogenes, on the forward (blue) or reverse (red) strand, within internal *var* gene clusters in the *Laverania*. Orientation is relative to the *P. falciparum* 3D7 reference genome.



Supplementary Fig. 12: DBLx is related to DBLε and the ancestral Duffy domain. To understand the diversity of DBLε and DBLn and to compare them to the newly described DBLx, a similarity matrix of all domains annotated as DBLε, Duffy, DBLpam1 and DBLx labelled sequences, see Supplementary Note 3. It can be seen that all domains are similar to each other and that the DBLx labeled sequences as defined by Larremore *et al*[31] cluster within a group that contains DBLε and Duffy domains (dotted box). *Pf, P. falciparum; Pprf, P. praefalciparum; Pr, P. reichenowi; Pbilc, P. billcollinsi; Pblac, P. blacklocki; Pgab, P. gaboni;* and *Padl, P. adleri*.

List of Supplementary Tables 1-9

see associated Excel file.

References

1 Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat Genet* **43**, 1031-1034, doi:10.1038/ng.937 (2011).

2 Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat Genet* **46**, 919-925, doi:10.1038/ng.3015 (2014).

3 Claessens, A. *et al.* Generation of antigenic diversity in Plasmodium falciparum by structured rearrangement of Var genes during mitosis. *PLoS Genet* **10**, e1004812, doi:10.1371/journal.pgen.1004812 (2014).

4 Ponnudurai, T. *et al.* Sporozoite load of mosquitoes infected with Plasmodium falciparum. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **83**, 67-70 (1989).

5 Mazier, D. *et al.* Complete development of hepatic stages of Plasmodium falciparum in vitro. *Science* **227**, 440-442 (1985).

6 Gerald, N., Mahajan, B. & Kumar, S. Mitosis in the human malaria parasite Plasmodium falciparum. *Eukaryotic cell* **10**, 474-482, doi:10.1128/ec.00314-10 (2011).

7 Nkhoma, S. C. *et al.* Population genetic correlates of declining transmission in a human pathogen. *Molecular ecology* **22**, 273-285, doi:10.1111/mec.12099 (2013).

8 Molineux, L. in *Malaria, Principles and Practice of Malariology* Vol. 2 (ed McGregor IA Wernsdorfer WH) 913-998 (London Churchill, Livingston, 1998).

9 Bopp, S. E. *et al.* Mitotic evolution of Plasmodium falciparum shows a stable core genome but recombination in antigen families. *PLoS Genet* **9**, e1003293, doi:10.1371/journal.pgen.1003293 (2013).

10 Udeinya, I. J., Graves, P. M., Carter, R., Aikawa, M. & Miller, L. H. Plasmodium falciparum: effect of time in continuous culture on binding to human endothelial cells and amelanotic melanoma cells. *Experimental parasitology* **56**, 207-214 (1983).

11 Payne, D. Spread of chloroquine resistance in Plasmodium falciparum. *Parasitol Today* **3**, 241-246 (1987).

12 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

13 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:btp698 [pii] 10.1093/bioinformatics/btp698 (2010).

14 Freedman, A. H. *et al.* Genome sequencing highlights the dynamic early history of dogs. *PLoS Genet* **10**, e1004016, doi:10.1371/journal.pgen.1004016 (2014).

15 Volkman, S. K. *et al.* Recent origin of Plasmodium falciparum from a single progenitor. *Science* **293**, 482-484, doi:10.1126/science.1059878 (2001).

16 Chang, H. H. *et al.* Malaria life cycle intensifies both natural selection and random genetic drift. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 20129-20134, doi:10.1073/pnas.1319857110 (2013).

17 Palstra, F. P. & Fraser, D. J. Effective/census population size ratio estimation: a compendium and appraisal. *Ecology and evolution* **2**, 2357-2365, doi:10.1002/ece3.329 (2012).

18 Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nature reviews. Genetics* **15**, 379-393, doi:10.1038/nrg3734 (2014).

19 Yasukochi, Y., Naka, I., Patarapotikul, J., Hananantachai, H. & Ohashi, J. Genetic evidence for contribution of human dispersal to the genetic diversity of EBA-175 in Plasmodium falciparum. *Malar J* **14**, 293, doi:10.1186/s12936-015-0820-2 (2015).

20 Castoe, T. A. *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8986-8991, doi:10.1073/pnas.0900233106 (2009).

21 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797, doi:10.1093/nar/gkh340 (2004).

22 Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* **27**, 221-224, doi:10.1093/molbev/msp259 (2010).

23 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564-577, doi:10.1080/10635150701472164 (2007).

24 Guindon, S., Delsuc, F., Dufayard, J. F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol Biol* **537**, 113-137, doi:10.1007/978-1-59745-251-9_6 (2009).

25 FigTree v.1.4.2, Available http://tree.bio.ed.ac.uk/software/figtree/ (2014).

26 Team, R. D. C. *R: A Language and Environment for Statistical Computing*. (2008).

27 Bastian, M., Heymann, S. & Jacomy, M. in *International AAAI Conference on Weblogs and Social Media* (2009).

28 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-1584 (2002).

29 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* **37**, W202-208, doi:10.1093/nar/gkp335 (2009).

30 Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431, doi:10.1186/1471-2105-11-431 (2010).

31 Larremore, D. B. *et al.* Ape parasite origins of human malaria virulence genes. *Nature communications* **6**, 8368, doi:10.1038/ncomms9368 (2015).

32 Wright, K. E. *et al.* Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. *Nature* **515**, 427-+, doi:10.1038/nature13715 (2014).

33 Guizetti, J., Barcons-Simon, A. & Scherf, A. Trans-acting GC-rich non-coding RNA at var expression site modulates gene counting in malaria parasite. *Nucleic Acids Res* **44**, 9710-9718, doi:10.1093/nar/gkw664 (2016).